Abstract:
Big Data deals with huge massive data that may be from heterogeneous autonomous sources. These data can be from distributed and decentralized control also. The management of huge data using traditional data management system is extremely challenging in day to day life. So it is mandatory to define new paradigm and re-evaluate current system to manage and process Big Data. In this paper, a review on some significant characteristics, issues and challenges related to Big Data management has been studied. Discussions of various Big Data Analytics framework that deal with Big Data analytics workloads have been discussed. Comparative study between the given frameworks and suitability of the same has been proposed.

Keywords: Big Data Analytics- Big Data Issues and Challenges- Apache Hadoop -Appache Drill-Project Storm.

1. Introduction

Digital world is flooded with huge amount of data generated by number of users worldwide. Diversity nature of data is due to as they come from different source and in various forms. To analyze a large volume of complex data, a specialized analytical platform data has been delivered relational database vendors.[1] This platform has different shapes, forms and different sizes.

The analytical services that run in third party hosted environments. In addition new technologies have emerged to address exploding volumes of complex data, including web traffic, social media content and machine generated data including sensor data, global positioning system data. For semi structured data to optimize ad-hoc queries, new non-relational database vendors combined with text indexing and natural language processing techniques with traditional database technologies has explored.

To analyze complex structured and unstructured data, a numbers of analytical platform are available in the market where each of which is designed to handle specific type of data/workload. In this paper, we have discussed three open source Big Data Analytics frameworks suitable for different types of workload.
2. Big Data Analytics characteristics

The term Big Data covers diverse technologies same as cloud computing. Web server logs, social networks, satellite imagery, traffic flow sensors; broadcast audio sensors, banking transaction etc are input to Big data Analytics[4].

Data should be analyzed in different dimension to recognize it as Big Data. Big Data can be characterized by different aspects. The commonly used aspects are volume, Velocity and Variety. Veracity and Value are main characteristics through which we can understand the nature of Big Data and the platform available to exploit them[2].

2.1. Volume: Huge amount of data generated by generated by different sources is increasing in petabytes and zetta bytes as infrastructure becomes increasingly available and affordable. This large amount of data is characterised as Big Data.

2.1. Velocity: Speed at which we are creating data is called as the sheer velocity and this is huge cause of Big Data[2]. Digital universe expands from 130 million to 40 trillion in 8 years (2005-2014). The data generated from various sources range include batch system to real time system. So this high velocity data defines new term called “Big Data”.

2.2 Variety: The diverse nature of representation of data generated by various sources data is big challenge in managing big data[2]. Data also take different forms like structured data for example ecommerce web sites deal with structured data, semi structured data like web server logs that deals with semi structured data and unstructured data like social websites that deals with unstructured data like audio, video, images etc[5]. The digital universe deals with combination of all structured, unstructured and semi structured data[2].

2.3. Veracity: Duo to sheer velocity of some data we cannot spend time in cleaning the data before using it. For making decision in business environment, the data should be precise and accurate. Hence combination of precise, imprecise, accurate, data can be called Big Data.

2.4. Value: New dimension of data called value obtained by processing huge volume, high velocity, variety and veracity of data[1]. Value represents the hidden knowledge represented by huge value of data. These hidden knowledge can be applied for business and getting competitive advantage from it represents value of Big Data.
3. Motivation for Big Data and Analytics

Statistics shows that there is an exponential increase in the rate of data generated on digital universe. Management and processing of such data cannot be handled effectively by current tools and Techniques. Extraction of hidden knowledge from these data cannot be done effectively by current tools. This knowledge which represents data is very important for decision making in business. When an enterprise can leverage all the information available with large data rather than just a subset of its data then it has a powerful advantage over the market competitors. As analysis of big data is very useful, some modified paradigms are essentials. Some of the application of Big Data in day to today life are given below.

3.1. Big Data Analytics and Health Care

Patient’s medical history, medication and personal details are stored and managed by Medical practitioners[8]. These are huge amount of data and cannot be managed by current tools. For example drug manufacturing company where the data that are very complex in nature and sometimes practitioners cannot correlated with other information, thus results in important information remains hidden. Some new advanced techniques for analyzing big data can be used and helps to extract hidden information which results in personalized medication. For example from the Advance analytics techniques can provide high insight into genetic and environmental causes of diseases.

3.2. Big Data Analytics and Intelligence Agencies

Satellite imagery, signal intercepts, and publicly available sources are used by intelligence agencies[8]. They collect very huge amount of data and connect all information to detect like thefts. Through this information threats, thefts and any other criminal activity can be prevented or detected. So very effective analytical techniques is required to deal with these data.

3.3. Big Data Analytics and Environment

Huge amount of data such as sensors monitoring air and water quality, meteological conditions, proportion of CO2, and other gas in air etc. air are used for understanding environment. The information like increased CO2 emission, increase or decrease of greenhouse effect can be found out from these huge amounts of data.

4. Big Data analytics issues and challenges

Organization dealing with Big Data is facing numerous challenges. Need of user and new technology has to understand by the system working with Big data[9]. Overcoming the challenges of Big data will be difficult[3]. Due to huge volume of data generated at different speed every day and variety of data expanding in high speed makes the task of handling it very difficult.
Current tools, technologies, architecture, management, and analysis approaches are unable to cope up with complexity of data presented. Some challenges are presented below.

**Privacy, Security and Trust** – Organization using Big Data, committed to protect the privacy, security of its users and should ensure that the organization must comply with all privacy and security related act to enhance the protection of and set clear boundaries for usage of personal information[9].

The increase in data and methods to control them should not affect the trust in the organization. The leakage of data or information into public domain will affect in the trust that users have in these agencies and their abilities to security hold information of a personal.

**Data Management and sharing** – Any data that the companies or agencies have, needs to be discoverable, accessible and usable. Even though agencies must should to try achieve these requirements but still adhering to privacy laws[9]. Open data makes relevant data available to public in the current era. Data that are open between companies should adhere to companies’ privacy laws of companies and it should be standardized to use it effectively and collaboratively between agencies. These points should be key focus of companies.

**Technology and Analytical skills** – Big Data and Analytics put lot of stress on ICT providers for developing new tools and technology to handle complex data. Current tools and technologies are unable to store, process and analyze massive amount of diverse data. Vendors and developers of Big Data systems and solutions including open source software are developing more capable tools to simplify the challenges of Big Data Analytics.

Some specific challenges related to Big Data and Analytics are:

**Data Storage and Retrieval** – Current storage can support managing data entry and data storage from small to huge. But handling very huge data like big data is not available[9]. The methods and techniques to handle semi or unstructured data for processing it is yet unknown.

**Quality Vs. Quantity** – When dealing with huge amount of data, sometime it is difficult to decide:

- Which data is inappropriate and how do we select most appropriate data?
- How do ensure authenticity of the data?
- How to estimate the value of data?

**Data Growth and Expansion**

As the organizations increases their services, their data is also expected to grow, Few organization also consider data expansion because of data grow in richness, data evolved with new techniques.
Speed and scale – When volume of data grows, it is difficult to gain insight into data within time period. Gaining insight into data is more important than processing complete set of data. Processing near real time data will always require processing interval in order to produce satisfactory output.

Structured and unstructured data – Transition between structured data stored in well defined tables and unstructured data (images, videos, text) required for analysis will affect end to end processing of data. Invention of new non-relational technologies will provide some flexibility in data representation and processing.

Data ownership – Very huge amount of data resided in the servers of social media service provider. These data is not really owned by them but they store data of their users. Actual owner of the page is one who has created the pager or account. This is ongoing and big challenge in area of social media.

Conclusion

In this work a detailed study of Big Data and Analytics has been performed. Following points related to Big Data and Analytics are worth noted. There is a requirement of Big Data Analytics frameworks for the organization that deal with different types of Big Data workloads. In addition middleware architecture is also required to integrate and process all Big Data related workloads. Organization dealing with Big Data and Analytics need to deal with challenges like privacy, security, data management and sharing, technology, skills and other specific challenges related to workload present in the organization.

References


**Corresponding Author:**

**Angulakshmi .M**

**Email:** angulakshmi.m@vit.ac.in